LANGUAGE FROM A COGNITIVE PERSPECTIVE

Grammar, Usage, and Processing

Studies in Honor of Thomas Wasow

*edited by*
Emily M. Bender & Jennifer E. Arnold
# Contents

1 Introduction 1  
**Jennifer E. Arnold & Emily M. Bender**

2 Grammar Engineering and Linguistic Hypothesis Testing: Computational Support for Complexity in Syntactic Analysis 5  
**Emily M. Bender, Dan Flickinger & Stephan Oepen**

3 Accuracy vs. Robustness in Grammar Engineering 31  
**Dan Flickinger**

4 Local Grammaticality in Syntactic Production 51  
**Ash Asudeh**

5 Blocking and the Architecture of Grammar 81  
**Peter Sells**

6 Simplicity and Fit in Grammatical Theory 99  
**Amy Perfors**

7 “Basic Information Structure” and “Academic Language”: An Approach to Discourse Analysis 121  
**James Paul Gee**

8 Relativizer Omission in Anglophone Caribbean Creoles, Appalachian, and African American
Vernacular English [AAVE], and Its Theoretical Implications  139
JOHN R. RICKFORD

9  Corpus-based Research on Language Production: Information Density and Reducible Subject Relatives  161
T. FLORIAN JAEGER

10  Ordering Choices in Production: For the Speaker or for the Listener?  199
JENNIFER E. ARNOLD

11  Weight and Word Order in Historical English  223
HARRY J. TILY

12  Investigating Syntactic Persistence in Corpora  247
NEAL SNIDER

13  Discontinuous Dependencies in Corpus Selections: Particle Verbs and Their Relevance for Current Issues in Language Processing  269
JOHN A. HAWKINS

14  Information in Virtual Spaces  291
SUSANNE RIEHEMANN
Corpus-based Research on Language Production: Information Density and Reducible Subject Relatives

T. Florian Jaeger

Preface
If it hadn’t been for Tom, I probably would have had to leave academia. Justified or not, in my fourth year of graduate school, I had grown disillusioned with certain aspects of theoretical linguistics. One of Tom’s classes stimulated my interest in psycholinguistics, and specifically what insights we can gain about the production system by studying the speakers’ choices in alternations. Tom and I started to collaborate on a project that soon grew into my thesis and further projects, including the one presented here. Tom is a fantastic advisor. I received guidance and support when I needed it without feeling pressured into working within any particular framework. He read countless versions of my thesis, papers, and talks, giving excellent feedback that often made me rethink my arguments. Tom always stayed open to arguments and was supportive even in those rare cases when we disagreed on how to proceed. Now that I am an advisor myself, I know to admire his dedication even more. Most probably, my own students would love for me to be more like him—alas, luckily, Rochester is far away from Stanford so that they can’t run away. Thank you, Tom, for your support and patience (e.g. when I was nowhere to be found until a few minutes...
before our joint talk in Cambridge...). With your help, it felt easy to start over and dive into psycholinguistics and cognitive science!

1 Introduction

When speakers encode their thoughts into linguistic utterances, there are often several ways to encode the same message. For example, speakers can encode messages into one or several clauses (e.g. *Move the triangle to the left of the square* vs. *Take the triangle. Move it to the left of the square*, cf. Gómez Gallo, Jaeger, & Smyth, 2008); within a clause, speakers can choose between different word orders (e.g. heavy NP shift, Arnold, Wasow, Losongco, & Ginstrom, 2000; Wasow, 1997; particle shift, Lohse, Hawkins, & Wasow, 2004; Wasow, 2002; the ditransitive alternation, Arnold et al., 2000; Bresnan, Cueni, Nikitina, & Baayen, 2007; Bock & Warren, 1985); speakers may morpho-syntactically contract certain words (e.g. *they are* vs. *they’re*, Frank & Jaeger, 2008) and may choose phonologically reduced variants for others (e.g. *strawdiny* vs. *extraordinary*, Johnson, 2004; *t/d*-deletion, and vowel weakening in English, Bell et al., 2003). In other words, choice points are present at all levels of linguistic production—all the way down to articulation (cf. variations in speech rate and articulatory detail, Aylett & Turk, 2004; Bell et al., 2003; Bell, Brenier, Gregory, Girand, & Jurafsky, 2009; Pluymaekers, Ernestus, & Baayen, 2005; Son, Beinum, & Pols, 1998; Son & Santen, 2005).

Much of Tom’s work has investigated choices during syntactic production, with a strong focus on word order variations (e.g. Wasow, 1997, 2002; summarized in Arnold, this volume). Here I present a study of a different type of syntactic choice, so-called syntactic reduction, where producers have a choice between omitting or producing optional function words. The study investigates syntactic reduction in passive subject-extracted relative clauses (henceforth **SRCs**), sometimes also called *whiz*-deletion. English **SRCs** can be produced in a full form with a relativizer and the auxiliary and a reduced form without these function words, as in (1b) and (1a), respectively.\(^1\)

---

\(^1\)I use the terms ‘reduced’ and ‘reduction’ without intending to imply a process where a full form undergoes change to become reduced. I merely refer to the fact that **SRCs** can be produced with or without the relativizer and auxiliary. Also, I use the term **SRC** reduction exclusively to refer to cases like (1). Some dialects of English also afford optional *that*-mentioning in **SRCs**. This type of alternation is not investigated here.
(1) The style of life . . .  
   a. . . . chosen by the beat generation . . .  
   b. . . . that was chosen by the beat generation . . .  
   is designed to enhance sexual experience.

The work presented here derived from collaborations with Tom that were inspired by a class he taught on Hawkins’s—as of then unpublished—book (Hawkins, 2004). Tom, Dave Orr (another student in Tom’s class), and I first started to investigate the processing mechanisms driving optional relativizer that-mentioning in non-subject-extracted relative clauses (e.g. *Most typos (that) you will find in this paper are unintentional*, see Jaeger & Wasow, 2006; Wasow, Jaeger, & Orr, in press; see also Fox & Thompson, 2007; Race & MacDonald, 2003). Tom noticed an intriguing pattern: Some apparently idiosyncratic findings reported in previous work (e.g. Fox & Thompson, 2007) were compatible with the hypothesis that more predictable relative clauses are less likely to have a relativizer. Inspired also by work on probability-sensitive phonetic reduction by Dan Jurafsky (Jurafsky, Bell, Gregory, & Raymond, 2001) and work on expectation-based comprehension by Roger Levy (Levy, 2005, 2008), we set out to investigate to ‘which extent’ speakers’ preferences in syntactic reduction are driven by redundancy. We hypothesized that speakers should be more likely to omit redundant material. If this hypothesis is correct, speakers should show a higher preference for reduced syntactic constituents if the constituent is predictable (redundancy is negatively correlated with predictability; see below). In several corpus studies on the reduction of non-subject-extracted relative clauses in spoken and written American English, we found that this is indeed the case: The more predictable a relative clause is given preceding words, the less likely speakers are to produce a relativizer (Jaeger, Levy, Wasow, & Orr, 2005; Jaeger, 2006; Levy & Jaeger, 2007; Wasow et al., in press; see also Jaeger, 2006, Ch. 3 on that-mentioning in complement clauses).

This hypothesis about the link between redundancy and syntactic reduction has since evolved into a theory of efficient language production. The hypothesis of Uniform Information Density (UID, Jaeger, 2006, 2010; Levy & Jaeger, 2007) states that producers prefer to distribute information uniformly across the linguistic signal. The information carried by a linguistic unit is defined in information theory as the logarithm-transformed inverse of its probability (Shannon, 1948). So,  

---

2 This one is for you, Tom. Thank you, for getting me started in the psycholinguistics world and all the insightful and enjoyable collaborations!
I(u) = \log \frac{1}{p(u)} = -\log p(u). That is, the less predictable something is, the more information it carries (and the less redundant it is). For the purpose of this paper, we can think of information density as the amount of information per word (but see Jaeger, 2010, for more detail).

Uniform information density can be shown to be a theoretically optimal solution for successful transmission of information across a noisy channel (cf. Aylett & Turk, 2004; Genzel & Charniak, 2002; based on Shannon, 1948). Thus, if there were no other constraints on language and language use, rational producers should distribute information absolutely uniformly across their linguistic signals. However, language is subject to many constraints and pressures, so this extreme case is unlikely to hold. For example, languages have to be learnable and their words and abstract structures need to be memorizable. But it is possible that a preference to distribute information uniformly is one of many pressures on language use. If such a preference affects syntactic production, we should observe that producers prefer syntactic variants that avoid peaks and troughs in information density. Choosing the full variant in syntactic reduction alternations, such as the SRC in (1) above, distributes the information that, for example, an SRC has started, \(-\log p(\text{SRC} |\text{context})\), over more words. Producers should then prefer the full variant (with that was) whenever the information at the SRC onset (e.g. on the verb chosen in (1) above) would otherwise be high.

This is the prediction of UID that I test in this paper using data from SRC reduction. Following much of Tom’s work, I take a corpus-based approach. I follow recent work (e.g. Bresnan et al., 2007; Jaeger, 2006, 2010) in using multilevel logit models to simultaneously control for other processing mechanisms known to affect syntactic reduction. Despite a rich tradition of influential corpus-based work (e.g. Arnold et al., 2000; Clark & Fox Tree, 2002; Clark & Wasow, 1998; Fox Tree & Clark, 1997; Hawkins, 2001, 2004; Resnik, 1996; Roland, Elman, & Ferreira, 2005; Wasow, 1997, 2002), corpus studies are arguably still under-represented in psycholinguistics. Unfamiliarity with the decisions involved in this type of research understandably makes it difficult to assess the results. In this paper, I sacrifice conciseness for a more detailed discussion of methodological issues.

2 Processing Accounts of Production Preferences

Existing processing accounts of preferences in language production can be grouped into four major categories. I introduce and compare them
with regard to their coverage across levels of linguistic production (see also Arnold, this volume; Hawkins, this volume).

Dependency processing accounts hypothesize that speakers prefer efficiently processable dependencies (Hawkins, 1994, 2001, 2004, 2007). Among other things, these accounts correctly predict that speakers prefer word orders that lead to shorter dependencies. For example, in word order variations following the verb (e.g. the ditransitive alternation), English speakers prefer to order long constituents after shorter constituents (Arnold et al., 2000; Arnold, Wasow, Asudeh, & Alrenga, 2004; Bresnan et al., 2007; Lohse et al., 2004; Wasow, 1997, 2002, inter alia). Crucially, speakers of verb-final languages show the opposite preference (for Japanese, see Yamashita & Chang, 2001; for Korean, Choi, 2007). Hence, across different languages, speakers seem to prefer word orders that order shorter constituents closer to the verb. Dependency processing accounts also receive support from syntactic reduction, such as that-mentioning (Elsness, 1984; Fox & Thompson, 2007; Hawkins, 2001; Jaeger, 2006; Race & MacDonald, 2003; Roland et al., 2005, inter alia). The prediction of dependency processing accounts for lower-level production choices, such as phoneme deletion and articulatory choices, are less clear. Hawkins discusses one such dependency processing account in more detail (Hawkins, 2004, this volume).

An alternative type of account holds that the relative ‘accessibility’ of referents is the primary driving force behind speakers’ preferences. There are two types of accessibility-based accounts: Alignment accounts and availability accounts. Alignment accounts (Bock & Warren, 1985; F. Ferreira, 1994) hold that speakers prefer to align grammatical function assignment with the relative accessibility of referents. Here accessibility refers to the conceptual accessibility of referents. A referent’s conceptual accessibility is affected by inherent properties (e.g. animacy, Bock, Loebell, & Morey, 1992; imageability, Bock & Warren, 1985), and by contextually conditioned properties (e.g. prior mention of the referent, Prat-Sala & Branigan, 2000).

Availability accounts (V. S. Ferreira, 1996; V. S. Ferreira & Dell, 2000; Levelt & Maassen, 1981; Prat-Sala & Branigan, 2000) also hold that the conceptual accessibility of referents affects speakers’ preferences at choice points in production. However, availability accounts consider the effect of accessibility to be more direct: According to availability accounts, speakers prefer to order accessible material early since it is available earlier for pronunciation (cf. the Principle of Immediate Mention in V. S. Ferreira & Dell, 2000). Here ‘accessibility’ refers to ease of retrieval from memory. Note that accessibility is not the same as constituent length or weight. A heavier constituent may have a more
accessible (and hence easier to retrieve) onset than a shorter constituent (cf. a very nice idea vs. xylophones). Since not all parts of a constituent have to be planned before speakers can initiate articulation (e.g. Griffin, 2003; Brown-Schmidt & Konopka, 2008), availability accounts do not necessarily predict that long constituents should be ordered after short constituents (contrary to Hawkins, 2007, 93-94). According to availability accounts, speakers choose between variants based on which one they can start pronouncing first (V. S. Ferreira & Dell, 2000, 289).

For word order variations in subject-initial languages, such as English, alignment and accessibility accounts make very similar predictions (for recent reviews, see Branigan, Pickering, & Tanaka, 2008; Jaeger & Norcliffe, 2009; see also Arnold, this volume). For syntactic reduction, such as the reduction of SRCs, the predictions of psycholinguistic alignment accounts are less clear. Availability accounts predict that speakers prefer to pronounce, rather than omit, optional words if the following material is not available for pronunciation. This prediction has received support from studies on that-mentioning in object-extracted relative clauses (Fox & Thompson, 2007; Jaeger & Wasow, 2006; Temperley, 2003) and complement clauses (Elsness, 1984; V. S. Ferreira & Dell, 2000; Roland et al., 2005), as well as from studies on SRC reduction (V. S. Ferreira & Dell, 2000, Experiment 3).

Availability accounts have been contrasted with ambiguity avoidance accounts (e.g. Bolinger, 1972; Hawkins, 2004; Temperley, 2003), which predict that speakers structure their utterances so as to avoid temporary ambiguities. Consider the SRC example in (2). Unlike in (1) above, the verb form chased is ambiguous out of context between the past tense and passive participle interpretation. In the reduced form (2a), this can cause comprehenders to be ‘garden pathed’ (Bever, 1970) into the unintended interpretation, where chased is temporarily interpreted as the past tense matrix verb to three mail men). Garden-pathing leads to comprehension difficulty at the disambiguation point (i.e. the word by). Speakers can avoid potential garden paths by producing a full SRC, as in (2b).

(2) Three mail men . . .
   a. . . . chased by dogs . . .
   b. . . . who are being chased by dogs . . .
   are indeed a hilarious sight.

There is evidence from corpus studies (Hawkins, 2004; Temperley, 2003) and experiments (Haywood, Pickering, & Branigan, 2005) that speakers sometimes avoid temporary syntactic ambiguities. However,
many recent corpus studies (Roland et al., 2005; Roland, Dick, & Elman, 2007) and experiments (Arnold et al., 2004; V. S. Ferreira & Dell, 2000; Kraljic & Brennan, 2005) have failed to detect ambiguity avoidance effects (see in particular, Roland et al., 2007 vs. Temperley, 2003). Whether speakers avoid syntactic ambiguity is therefore still very much an open question.

3 The Production of Subject-Extracted Relative Clauses

Consider example (3). The information carried by the SRC onset, left, differs between the full and the reduced variant. In the reduced variant, (3a), left signals the beginning of an SRC constituent. This is in addition to its lexical information. In the full SRC variant, (3b), however, left carries less information, since the preceding words (that was) already signal the presence of an SRC constituent.³

(3) More work on government projects helped close the gap . . .
   a. . . . left by cash-strapped home-owners.
   b. . . . that was left by cash-strapped home-owners.

In information theoretic terms (Shannon, 1948), the information carried by left can be described as the sum of two pieces of information: The information that there is an SRC, I(SRC | ctxt), and the information that the first word in that SRC is left, I(left | ctxt, SRC). Put more generally:

\[
I(SRC \text{ onset} | \text{ ctxt}) = I(SRC | \text{ ctxt}) + I(onset | \text{ ctxt, SRC})
\]

If present, the optional function words spread the same information over two more words, thereby lowering information density. The function words reduced the information carried by the first content word in the SRC. To illustrate this, consider the extreme case where the function words signal an SRC with 100% certainty. In that case, the function words carry the information that there is an SRC (the left half of Equation 1), and the onset carries only its lexical information (the right half of Equation 1). If, as predicted by UID, producers avoid peaks and troughs in information density, they should prefer the full variant for SRC onsets that are high in information content. Given Equation 1,

³The two function words carry additional information, e.g. the tense of the SRC. Here I simplify and focus on the constituent boundary information.
producers should be more likely to use the full form, (a) the less probable an SRC is given the preceding context and (b) the less probable the words at the SRC onset are given the preceding discourse and given that there is an SRC. Here, I focus on the first prediction: The more predictable an SRC, the more likely producers should be to use the reduced form without the optional function words.

Like any methodology, the corpus-based approach involves judgment calls. While some of these are standardly part of the write-up, others are usually not mentioned for the sake of brevity. Here I am more explicit about the process that leads from the hypothesis formulation via a priori considerations (such as choice of model, choice of control predictors, etc.), corpus selection, data extraction, data exclusion, and variable preparation to the statistical analysis. I address general considerations and the specific decisions made for the current study (though the list of issues addressed below is by no means exhaustive).

3.1 A Priori Consideration: Envisioning the Analysis

Before embarking on a corpus-based project, it is a good idea to look ahead and to envision the type of analysis that will be necessary. While it is almost inevitable that inspection of extracted data will force revisions of initial assumptions, there are a few considerations that should guide decisions from corpus selection to statistical analysis. For example, what exactly is the outcome (dependent variable) of interest and what type of data should hence be collected? What predictors will be included in the analysis to avoid confounds and to model the predictions of alternative accounts? And approximately how many parameters will be required? Based on the anticipated number of parameters and the type of analysis required, how much data is necessary to have a chance to detect the hypothesized effects? I address these questions in the order mentioned.

Outcome of Interest

It is important to understand the outcome of interest. For example, for corpus-based studies on relativizer that-mentioning in Standard American English, only finite, restrictive, non-pied-piped, non-extraposed, non-subject-extracted relative clauses should be included in the analysis, because only those types of relative clauses allow both the full and the reduced form (cf. Jaeger, 2006, Appendix C). In other words, when we study producers’ preferences in production, only cases where producers actually have a choice between at least two variants should be included. Failure to do so is likely to lead to spurious results. For the current study, only passive subject-extracted relative clauses are of
interest, since active subject-extracted relative clauses do not afford the same type of choice (*whiz*-deletion, as in (1) above). Additional annotation (see below) excluded further cases that were incompatible with the alternation.

**Likely Predictors**
Whenever possible, predictor selection should be driven by *a priori* theoretical considerations (such as the need to account for theories of sentence production) as well as empirical considerations (such as what factors are known to affect the outcome).

Based on previous research on SRC reduction (V. S. Ferreira & Dell, 2000; Hare, Tanenhaus, & McRae, 2007; Trueswell, 1996) and other reduction phenomena (see references in Section 2), there are several controls that should be included in any serious test of a new hypothesis that makes predictions about SRC reduction. Since different corpora provide different annotations, it is worth considering what types of information will be needed for the model before choosing a corpus for the study. For the current purpose, I include the following controls (though additional controls may be necessary, as I discuss in Section 3.6):

- **length of SRC** to account for dependency length effects (cf. Elsness, 1984; Fox & Thompson, 2007; Hawkins, 2001; Jaeger, 2006, 2010; Lohse, 2000; Quirk, 1957; Race & MacDonald, 2003; Roland et al., 2005),
- **word form frequencies** of the word immediately preceding the SRC as well as the frequency of the first word in it to control for availability-based effects on sentence production (cf. V. S. Ferreira & Dell, 2000; Jaeger & Wasow, 2006; Race & MacDonald, 2003),
- whether the participle is potentially ambiguous, since producers may avoid ambiguity (cf. Bolinger, 1972; Temperley, 2003; Hawkins, 2004).

Assuming that the SRC length effect may require modeling of non-linearities (as is the case for dependency length effects on complementizer-mentioning, Jaeger, 2006, 2010), a rule-of-thumb estimate of the number of required control parameters would be about 9 (one parameter for ambiguity, one each for the two frequency measures, 3 each for SRC length, plus 50% for predictors not considered *a priori*). Adding one parameter each for two estimates of the information density at the SRC onset (information density can be estimated given preceding material and given the words at the SRC onset, see Section 3.4 below), this comes to an estimate of 11 parameters. This lower bound estimate can
be used for considerations regarding the avoidance of overfitting, which I describe next.

**Avoiding Overfitting: How Much Data will be Necessary?**

Once the approximate number of parameters for a model is known, it is possible to estimate how much data (i.e. how many observations) will be necessary to avoid overfitting. An overfitted model describes the sample rather than generalizing beyond the sample to the population. Exactly how many observations are necessary to avoid overfitting depends on the distribution of the dependent and independent variables and correlations between predictors that one wishes to distinguish between. Hence, there is no simple rule that is guaranteed to avoid overfitting. However, several rules of thumb have been suggested. Generally, 10-15 times as many observations as numbers of parameters in the model seems to be sufficient (for a very readable introduction to considerations about overfitting, see Babyak, 2004; for power simulations assuming effect sizes that are typical for behavioral research, see Green, 1991). Crucially, it is the limiting sampling size that matters. So, for binary outcomes like the choice of full over reduced passive SRC, it is the minimum number of observations of the less frequent outcome that should be 10-15 times higher than the number of parameters in the model (for simulations, see Peduzzi, Concato, Kemper, Holford, & Feinstein, 1996).

These rules of thumb for number of necessary observations have to be taken with caution. Especially for corpus-based work, they should be treated as a lower bound. It helps to consider a standard psycholinguistic production experiment. For example, V. S. Ferreira and Dell (2000, Experiment 3) present a recall experiment on passive SRCs with 48 subjects and 24 items for a 2 x 2 x 2 design (i.e. 8 parameters including the intercept). About 10% of the 1,152 data points had to be excluded from the analysis for a variety of reasons. Participants produced full SRCs in about 30% of all trials. Hence, the less frequent outcome occurred a bit over 300 times, resulting in an approximately 4 times higher ratio of data to parameters than suggested above as the absolute lower bound (8 * 10 = 80; most recall experiments contain 4-6 items per condition and thus yield even more statistical power than the experiment described here).

Corpus data are usually highly unbalanced, which further reduces power since it leads to many cells in the overall design that contain little to no data. Hence, I recommend a more conservative criterion: For highly unbalanced data and complex analyses with many parameters, the less frequent outcome should be about 50-100 times more frequent
than the number of parameters in the model. It is important to keep in
mind though that these are rules of thumb. Several other factors deter-
mine whether an effect can be detected (e.g. effect size, distribution of
the outcome with respect to the predictor, and the presence of collinear
predictors that need to be accounted for).

A comparison of corpus analyses with experimental work on syn-
tactic reduction suggests that experiments generally overestimate the
relative frequency of full over reduced syntactic variants (cf. exper-
iments in V. S. Ferreira & Dell, 2000; V. S. Ferreira, 2003 vs. corpus
data in Roland et al., 2005; Jaeger, 2010). Hence, full SRCs are likely
to be the less frequent outcome, possibly making up about 10-20% of
all SRCs. Assuming this rate of full SRCs and assuming the number of
parameters needed for the model (see previous section), a corpus with
at least 550 to 1,100 or preferably closer to 5,500 to 11,000 SRCs is
needed for the current study!

3.2 Selecting a Corpus

Now that we have an estimate of the lower bound for the number of
observations needed for the current study, the next step is to find a
suitable corpus. Several considerations affect what corpus is best suited
for a research project. Corpora differ in terms of the type of language
(e.g. in terms of modality, register, style, genre, monologue vs. dialogue
data), available annotation (e.g. syntactic annotation vs. plain text),
and the distribution of the event of interest. Available annotation af-
ffects how much time and money will have to be spent on additional
annotations before it will be possible to extract all necessary informa-
tion about the outcome and the predictors. Corpus size and distribution
of the event of interest affect how much power the analysis will yield.

Speech corpora are arguably more appropriate for research on lan-
guage production. However, passive SRCs are very infrequent in speech.
For example, less than 500 passive SRCs are observed in the Switchboard
corpus of American English telephone dialogues (Godfrey, Holliman, &
McDaniel, 1992) of approximately 800,000 spoken words. Preliminary
tests suggest that as few as 10-20% of these passive SRCs in these cor-
pora are in the full form. Given the low-frequency of full SRCs, this
suggests that even the 10 million spoken words of the British National
Corpus would not provide enough data for the current study. When the
most suited corpora do not contain the outcome of interest in sufficient
numbers, researchers are faced with the decision of whether other cor-
pora are appropriate approximations. Alternatively, the target struc-
ture can be elicited in controlled experiments.
Here, I take the former approach. The written portion of the British National Corpus (BNC) is used as a data source for the current study. This decision was made because there is some evidence that written language by and large exhibits the same processing pressures observed in speech (Jaeger & Wasow, 2005). More specifically, information density effects on that-mentioning have been observed in writing (Jaeger et al., 2005). The written parts of the BNC contain approximately 90 million words from various genres and styles (see Nation, 2004, 11). The corpus was automatically parsed by Benjamin Van Durme using the Charniak-Johnson parser (Charniak & Johnson, 2005). Additionally, all sentences were annotated with author and genre information using BNC meta data and scripts provided by Judith Degen.

3.3 Corpus, Data Extraction, Annotation, and Exclusion

After a corpus is selected, the relevant cases need to be extracted, along with the outcome and predictor values. Before the data can be entered into the analysis, erroneously included cases and outliers, as well as cases with missing or unreliable information, need to be excluded. Each of these steps includes decisions that can affect the final analysis and should therefore be handled with care. If a corpus is not already annotated for all variables of interest, additional manual annotation may be required, which requires the researcher to make further judgment calls.

Data Extraction

The syntactic search software TGrep2 (Rohde, 2005) was used to extract all passive SRCs of a certain type from the corpus. Whenever data is automatically extracted from corpora, there is a danger of introducing biases into the analysis by either erroneously in- or excluding cases. This danger is especially high when, as in the present case, the data are extracted by means of syntactic searches over an automatically parsed corpus that has not undergone manual correction by annotators. Blind reliance on the consistency of automatic syntactic annotation risks in- or excluding cases in a non-random and yet uncontrolled fashion. Statistical analysis of such data may simply reflect whatever features the parser is conditioned on and hence lead us to report spurious results as theoretically relevant. There are several strategies to reduce such confounds (see also Roland et al., 2007, Appendix A). If the outcome of interest is highly frequent, so that a small corpus can be used, the entire corpus can be manually annotated. If, as in the present case, a manual syntactic annotation of the entire corpus is not feasible, automatic syntactic searches have to be used in such a way that they do not introduce a bias.
To avoid biases, the first step is to become familiar with the annotation guidelines used for the corpus. For the current case, almost all full SRCs are marked as S modifying an NP, where as almost all reduced SRCs are marked as VP modifying an NP, as in Figure 1a and Figure 1b, respectively.

Unfortunately, there is considerable variation in how exactly the S or VP attaches to the NP it modifies and how the internal structure of the S or VP was annotated. Additionally, exploratory searches revealed that the participle (bought in Figure 1) is not always marked as such (VBN in Penn Treebank annotation)—sometimes it is misparsed as past tense verb (VBD) or as adjective (JJ). All in all, SRC parses are relatively unreliable. For that reason, I employed search patterns that mostly rely on linear order information and only make reference to structural relations that are parsed reliably. Since false exclusions are harder to detect than false inclusions, I started with a very inclusive pattern based on only linear order constraints. For feasibility, only SRCs immediately following a noun are considered. This drastically reduces the number of false inclusions while not introducing a bias (UID predictions hold both for SRCs that are adjacent and SRCs that are non-adjacent to the noun). Hence, the first search looked for nouns that were either immediately
adjacent to a word form compatible with a participle interpretation or only separated from it by a valid relativizer-auxiliary combination (e.g. *which had been, who are, or that will be*). Instead of relying on correct part-of-speech tagging for the participles, a regular expression matched any adjective ending in */-ed/ as well as all past tense and participle forms.

This pattern yields a very large number of hits (> 100,000), but contains many false inclusions (estimated by checking a small random sample). I then successively refined the pattern. At each step, I checked that only falsely included cases were excluded by the refined pattern (using the same random sample). One criterion that considerably reduced the rate of false inclusions, while not leading to false exclusions, was the requirement for a *by-PP* immediately following the participle:

\[(4) \ldots \text{Noun (Rel Aux) Participle by-PP} \ldots \]
\[
\ldots \text{cars (that are) bought by companies} \ldots
\]

This excludes many false inclusions where a matrix verb is mistakenly parsed as a reduced passive SRC. While limiting matches to SRCs with *by-PPs* is non-random, it crucially does not introduce a bias into the analysis: UID’s predictions are the same for SRCs with and SRCs without PPs. This decision is comparable to the decision of an experimenter to use only SRC stimuli with *by-PPs*.

Successive refinement led to the final TGrep2 search pattern, shown in Table 1. The syntactic tree of an example match is given in Figure 2. This pattern yielded 50,077 matches.

\[
*=\text{srcverb} , (\text{/NN/=nhead} \gg (\text{/NP/} < (*=\text{src} \ll (=\text{srcverb} !<< (*=\text{srcverb} \ll (=\text{nhead})))})
\]
\[
. (by \gg (\ast \gg /\text{PP}/))
[= /\text{VB}(N|D)/
|\ast /\text{JJ}/ \ll ./+.ed$/
|!$ /\text{NP}[NN\]})}
\]

**Table 1** TGrep2 search pattern for reduced SRCs

The result returned by the final pattern still contains a considerable number of erroneous inclusions. It would be possible to stop at this point and deem the error rate acceptable. This is an efficient approach as long as the remaining in- and exclusion errors are distributed non-randomly with regard to the outcome (for an example of such an approach, see Roland et al., 2005). Here, a different approach was taken. I first applied several automatic filters to exclude data that should definitely not be included. Just as subjects may skip a case or a device
FIGURE 2  Tree representation of an example match for the SRC TGrep2 search pattern

may have a measurement error in experiments, data extracted from a corpus may miss variable information. For example, many corpora contain annotation mistakes or inconsistencies that make it impossible to extract all information about all cases. Especially when there is a lot of data, it is easier to simply exclude those cases than to manually annotate the missing information (but caution is necessary, if there is any reason to believe that missing information is not randomly distributed over cases). This left over 48,000 cases (see Table 2 below for details).

Given the large number of remaining cases, I decided to further exclude cases with infrequent idiosyncratic properties that may or may not influence processing—a process I refer to as ‘stratification’. At each step, both full and reduced SRCs not meeting the inclusion criteria were excluded. Cases for which variable information could not be estimated with high reliability were also excluded.

After all automatic filters were applied, a manual annotation was conducted on the remaining data. The purpose of the annotation was to further reduce the rate of erroneous inclusions and to check how accurately the values for the predictors were extracted. Next, I describe the different data exclusion steps and the annotation.⁴

Data Exclusion: Stratification vs. Control

Data sampled from corpora are often highly unbalanced with regard to infrequent effects. Unfortunately, infrequent effects can nevertheless

⁴The information density effect reported below was stable through all exclusion criteria, as well as additional tests. The additional annotation did not change the results either, although it was conceivable that it would.
be large. Consider adjacency effects on *that*-mentioning. Relative and complement clauses that are not adjacent to their head exhibit *substantially* higher rates of *that*-mentioning than adjacent clauses (Bolinger, 1972; Elness, 1984; Fox & Thompson, 2007; Hawkins, 2001; Race & MacDonald, 2003). Since the effects are strong, it seems advisable to include corresponding controls in the model. However, non-adjacent complement or relative clauses are relatively rare (e.g., < 5% in spontaneous speech, Jaeger, 2006, 57 & 104). In such cases, it can be easier to exclude all of the infrequent cases (stratification) rather than to lose degrees of freedom in order to account for the effect. The main advantage of stratification is peace of mind: Not only is the potential effect held constant, but all of its potential interactions with other predictors in the model are held constant, too. By avoiding more control parameters, we also decrease the chance of overfitting. Fewer parameters also mean more power, although the decision to exclude data may also decrease power. So it’s important to carefully consider the trade-off between fewer parameters and less data (e.g., by taking into consideration the rule-of-thumb mentioned above, according to which the minimum number of observations of the less frequent outcome should be 10-15 times higher than the number of parameters). Here, extremely long SRCs (> 23 words, which is 2.5 standard deviations above the mean) were excluded.

If applied appropriately, such stratification can be an effective way to reduce unaccounted-for variance in the data while keeping enough data for the final analysis. Laboratory experiments can be thought of as employing extreme stratification. Most psycholinguistic experiments use small sets of stimuli types. Experiments either hold properties that are known to affect the outcome constant across conditions, or they hold the mean of those properties constant (the latter is more common, but can be insufficient, see Baayen, 2004).

Stratification can also be a good solution when values of predictors cannot be reliably estimated. For example, for the present study, it is necessary to estimate the information content of the SRC onset. These estimates are based on conditional probability estimates (see Section 3.4 below), and conditional estimates are unreliable if the value being conditioned on is infrequent. Consider an extreme case. If the word *snark* occurs exactly once in the BNC and that one time it is followed by an SRC, would we want to conclude that SRCs are entirely predictable

---

5 Data exclusion has to be conducted on a priori defined criteria, such as power considerations described above. If multiple tests are conducted on different subsets of the data, the family-wise error rate will be inflated and appropriate corrections have to be applied.
after the word *snark*? Intuitively, the situation would be different if we had seen the word *snark* 100 times, each time followed by an SRC. Here cases were excluded if the head noun or the participle occurred less than 50 times in the corpus since the information density estimates introduced below are conditioned on the head noun and the participle. Note that this does not a priori exclude highly information dense SRC onsets from the data, since the conditional probability of an SRC is logically independent of the absolute frequency of the event conditioned on.

Cases with an SRC participle that was observed less than 5 times in the database were also excluded. This improves control for verb-specific effects in the rate of syntactic reduction (included in the analysis as random effect, see below). The effect of all automatic exclusions is summarized in Table 2.

<table>
<thead>
<tr>
<th>Exclusion criterion</th>
<th>Cases</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Automatic Detection of Inclusion Errors</strong></td>
<td>1,905</td>
<td>4.0%</td>
</tr>
<tr>
<td>Wrong parse (e.g. idioms)</td>
<td>25</td>
<td>0.1%</td>
</tr>
<tr>
<td>Missing variable information</td>
<td>1,407</td>
<td>2.9%</td>
</tr>
<tr>
<td>Match for participle not a verb</td>
<td>338</td>
<td>0.7%</td>
</tr>
<tr>
<td><strong>Stratification</strong></td>
<td>8,931</td>
<td>17.8%</td>
</tr>
<tr>
<td>SRC length in words &gt; 23</td>
<td>3,392</td>
<td>6.8%</td>
</tr>
<tr>
<td>Frequency of preceding noun form in corpus &lt; 50</td>
<td>1,606</td>
<td>3.2%</td>
</tr>
<tr>
<td>Frequency of participle form in corpus &lt; 50</td>
<td>490</td>
<td>1.0%</td>
</tr>
<tr>
<td>Frequency of participle in database &lt; 5</td>
<td>5,151</td>
<td>10.3%</td>
</tr>
<tr>
<td>Remaining cases</td>
<td>39,241</td>
<td>78.4%</td>
</tr>
</tbody>
</table>

**TABLE 2** Summary of all automatic exclusion criteria and number of cases that did not meet criterion. Bold lines summarize the exclusions listed below them. The last column relates the exclusions to the number of hits returned by the search pattern in Table 1. The total number of exclusions is less than the sum of the exclusions since many cases failed to meet several conditions.

**Data Annotation**

Next, I discuss the manual annotation that was added in order to further exclude erroneously included cases and correct values of input variables. Three annotators (see acknowledgments) annotated 18,305 of the remaining cases in time for this article.

Annotators determined whether cases were actually passive SRCs and, if so, whether they were compatible with both the full and reduced form. Annotators also marked cases for which the entire sentence could not be understood (the BNC contains headlines and other sentence fragments). Those were later excluded from the analysis.
Additionally, annotators checked \texttt{SRC} length information since extraction of \texttt{SRC} length based on the automatic parses turned out to be unreliable. Automatically identifying the end of a relative clause is much harder than identifying where it starts. Even for human annotators, and sometimes even with additional context, it is hard or impossible to determine where an \texttt{SRC} ends for cases like (5), where the italicized material could modify the \textit{the site}, in which case it is not part of the \texttt{SRC}, or it could modify the \texttt{SRC}'s verb phrase. Such cases were marked and excluded from the analysis.

(5) This was the site chosen by the jesuits \textit{for their complex of college and churches} . . . .

To avoid degradation of the bigram estimates of information density used below, annotators also marked cases in which the modified NP was complex (e.g. conjoined NPs, partitive NPs), where the actual head noun was a multi-word compound, or where the actual head noun was not adjacent to the \texttt{SRC}.

The analysis was based on only \texttt{SRC}s that were (a) judged to be compatible with both the full and the reduced forms, (b) adjacent to the head noun of the modified NP, and for which (c) all variable information could be determined. The exclusions based on the manual annotation are summarized in Table 3.

<table>
<thead>
<tr>
<th>Exclusion criterion</th>
<th>Cases</th>
<th>% of Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence incomprehensible</td>
<td>568</td>
<td>3.2%</td>
</tr>
<tr>
<td>\texttt{SRC} length could not be determined reliably</td>
<td>1,016</td>
<td>5.5%</td>
</tr>
<tr>
<td>Not an \texttt{SRC}</td>
<td>434</td>
<td>2.4%</td>
</tr>
<tr>
<td>\texttt{SRC} judged to be non-alternating</td>
<td>323</td>
<td>1.7%</td>
</tr>
<tr>
<td>Complex modified NP</td>
<td>1,767</td>
<td>9.7%</td>
</tr>
<tr>
<td>Multi-word head noun of modified NP</td>
<td>16</td>
<td>0.1%</td>
</tr>
<tr>
<td>Head noun of modified NPs non-adjacent to \texttt{SRC}</td>
<td>0</td>
<td>0.0%</td>
</tr>
<tr>
<td><strong>Remaining cases</strong></td>
<td>13,596</td>
<td>74.3%</td>
</tr>
</tbody>
</table>

TABLE 3 Summary of all exclusion criteria and number of cases that did not meet criterion. The last column relates the exclusions to the total number of \textit{annotated} cases. The total number of exclusions is less than the sum of the exclusions since many cases failed to meet several conditions.

This left 13,596 \texttt{SRC}s, of which 899 (6.7\%) were full \texttt{SRC}s. These \texttt{SRC}s contain 546 different participle forms, of which only 31 occur at least 100 times in the database. Figure 3 summarizes the distribution of participles types in the database and lists the 50 most frequent participles, which together account for over 61\% of the data.
3.4 Variable Preparation and Analysis

Next, I describe the type of regression model used to analyze the data, summarize the predictors in the model, and then describe the most common issues with this type of analysis.

**Type of Analysis: Mixed Logit Regression**

A mixed logit model (Breslow & Clayton, 1993) was used to analyze the effect of information density predicted by UID on producers’ choice of full over reduced SRCs, while controlling for effects predicted by alternative accounts. The model was fit using the function `lmer` from the `lme4` library (Bates, Maechler, & Dai, 2008) within the statistics software package R (R Development Core Team, 2008). A detailed introduction to mixed logit models is beyond the scope of this paper. I briefly summarize their main properties (for a technical introduction, see Agresti, 2002; Gelman & Hill, 2006; for introductions intended for psycholinguistic audiences, see Baayen, 2008; Jaeger, 2008).

Mixed logit models can be thought of as an extension of logistic regression. They are intended for the analysis of binomially distributed categorical outcome variables. The outcome (a.k.a. dependent variable) is regressed against the predictors (a.k.a. independent variables).
Logistic regression assumes that all observations were sampled randomly and hence are independent of each other. However, this assumption is typically violated in psycholinguistic research. The database for the present study, too, contains violations of independence. For example, many producers contribute more than one observation. Since producers may have different base rates of SRC reduction, this can make the analysis unreliable. Intuitively, we can learn more about how SRCs are distributed in English when we have 200 data points each from 20 speakers than when we have 4,000 data points from one single speaker. Analyses that fail to adequately account for violations of independence generally overestimate the amount of data and hence the confidence one can have in the results. Unlike ordinary logistic regression, mixed logit models can restore the assumption of (conditional) independence. Mixed logit models allow researchers to specify random effects that efficiently account for, e.g., random differences between producers (for an excellent introduction see Gelman & Hill, 2006). Here, two random effects were included: One to account for differences between writers and one to model idiosyncratic preferences of different participle forms.

Predictors in the Model

Table 4 summarizes all fixed effect predictors in the model along with the number of parameters used to model them. I first describe the control predictors and then the two information density measures.

The length of the SRC in words excluding the omissible function words (MIN= 3, MAX= 23, MEAN= 7.6, STDEV= 4.5) was included to control for effects of dependency processing (Hawkins, 1994, 2004). SRC length was modeled with restricted cubic splines (Harrell, 2001, 16-24) since previous work had found non-linear effect of constituent length on syntactic reduction (Jaeger, 2006, 2010).

The log-transformed frequency of the first word in the SRC (MIN= 4.2, MAX= 12.0, MEAN= 8.8, STDEV= 1.3), the participle, was included to control for effects of availability-based production (V. S. Ferreira & Dell, 2000; Levelt & Maassen, 1981; Jaeger & Wasow, 2006; Race & MacDonald, 2003). The log-transformed frequency of the head noun of the modified NP (MIN= 3.9, MAX= 15.5, MEAN= 8.7, STDEV= 2.0) was included to capture potential ‘spill-over’ effect due to high processing load immediately preceding the SRC onset.

To account for potential ambiguity avoidance effects, a predictor coding whether the participle was ambiguous was included in the model. This corresponds to what Section 3.3 Wasow and Arnold (2003) call ‘syntactic ambiguity’. Under this coding, most SRCs in the database (93%) have a potentially ambiguous onset. However, many—if not
The name and description of each input variable are listed; the last column describes the predictor type (’cat’ = categorical, ’cont’ = continuous) along with the number of parameters associated with the predictor.

most—syntactically ambiguous SRCs are not actually ambiguous in context. That is, many of the cases labeled as potentially ambiguous here are unlikely to cause garden path effects. In Wasow and Arnold’s terms, they are not ‘pragmatically ambiguous’. Unfortunately, only syntactic ambiguity annotation was available for the current study. I return to this issue below.

Finally, two measures of the information content of the SRC constituent boundary were included in the model. The information content given the preceding context was estimated using a simple bigram, $I(\text{SRC} \mid \text{NOUN}) = -\log p(\text{SRC} \mid \text{NOUN})$ (MIN= 2.1, MAX= 11.8, MEAN= 5.7, STDEV= 1.2). Additionally, the information content of the SRC onset was also estimated given the participle, $I(\text{SRC} \mid \text{PARTICIPLE}) = -\log p(\text{SRC} \mid \text{PARTICIPLE})$ (MIN= 1.2, MAX= 9.4, MEAN= 3.3, STDEV= 1.0). This can be understood as an approximation of how much information the presence of an SRC carries after the participle has been processed.

### Preparing Variables for Analysis: Common Issues

As regression models allow us to assess the partial effects of multiple predictors, they allow the simultaneous test of multiple hypotheses in one model. Like for any statistical procedure, what specific hypothesis is being tested depends on the precise way the input variables were

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Description</th>
<th>Type(βs)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Intercept</strong></td>
<td></td>
<td>(1)</td>
</tr>
<tr>
<td><strong>Domain length</strong></td>
<td>Length of SRC</td>
<td>cont(3)</td>
</tr>
<tr>
<td><strong>Lexical retrieval at SRC onset</strong></td>
<td>Log frequency of participle</td>
<td>cont(1)</td>
</tr>
<tr>
<td><strong>Lexical retrieval before SRC onset</strong></td>
<td>Log frequency of head noun</td>
<td>cont(1)</td>
</tr>
<tr>
<td><strong>Potential ambiguity at SRC onset</strong></td>
<td>SRC onset potentially ambiguous w/o optional function word</td>
<td>cat(1)</td>
</tr>
<tr>
<td><strong>Information content of SRC onset</strong></td>
<td>SRC onset’s information content given preceding word</td>
<td>cont(1)</td>
</tr>
<tr>
<td></td>
<td>SRC onset’s information content given its first word</td>
<td>cont(1)</td>
</tr>
<tr>
<td><strong>Total number of parameters in model</strong></td>
<td></td>
<td>9</td>
</tr>
</tbody>
</table>

TABLE 4
entered into the model. It is useful to distinguish between the input variables to a regression and the actual predictors in the model. Several decision steps lie between the choice of input variables and the final model. Unfortunately, a comprehensive description of the steps involved in regression analysis for corpus-based research is beyond the scope of this paper (for a tutorial, see Jaeger & Kuperman, 2009; for introductions to regression modeling, see Baayen, 2008; Gelman & Hill, 2006).

Figure 4 provides an overview of the process I followed in preparing and entering the predictors described above into the analysis.

![Figure 4](image_url)

**FIGURE 4** Visualization of the process from variable selection to model evaluation involved in regression analyses.

Initial data analysis leads to the input variables. It is a good idea to have at least an impression of the distributions of the predictor and outcome variables. The input variables are then transformed (dashed lines in Figure 4 indicate that the variable changes; the labels above the dashed lines give an example of what type of operation may be applied to the variable), as well as coded and/or centered and possibly standardized, before we create higher order terms based on any of the predictors (e.g. interactions or non-linear terms for continuous predictors, such as the restricted cubic splines for SRC length). At several steps during this process, we should check for outliers since they...
can be overly influential. During the initial data exploration, we may exclude cases that must clearly be measurement errors or that miss variable information. After the predictors have been transformed (cf. log-transformation of the frequency measures above), it is possible to exclude outliers based on distributional assumptions.

Collinearity must be assessed in the model to provide reliably interpretable coefficients. Collinearity can lead to biased standard error estimates. The good news is that collinearity increases standard error estimates (and hence Type II errors) for only those predictors in the model that are collinear. If there is collinearity in the model for the predictor of interest there are a variety of strategies to reduce it, such as centering, stratification, residualization, principal component analysis, and so on (see Baayen, 2008). Here all predictors were centered. Residuals were used to remove collinearity between the information density measures and the log-transformed frequency of the modified NP’s head noun and the SRC participle. The log-transformed frequency of the head noun of the modified NP is highly correlated with the information density estimate based on that head noun ($r = 0.61$). The SRC onset’s information content based on the head noun was regressed against the head noun’s log-transformed frequency and the residuals were entered into the model instead of the original information density estimate. Although the correlation between the information density estimate based on the participle and the participle’s log-transformed frequency was comparatively small ($r = 0.31$), the same residualization procedure was applied. The residualized information density estimates capture the proportion of information density that is not explicable by the log-transformed frequencies of the participle or the modified NP’s head noun. No other predictors needed to be residualized.

3.5 Results

The final step before we can interpret the model is model evaluation. If model evaluation suggests that the model is flawed, some of the steps outlined above may need to be repeated.

Model Evaluation

The model with the information density predictors and all controls is significantly better than the baseline model (always guessing the more probable outcome given only random effects). A comparison between the model and the baseline model with only the random effects of writer and participle shows that the fixed effect parameters in the model are justified ($\chi^2_{\Delta(\Lambda)}(8) = 329.5, p \ll 0.0001$). However, the model also shows need for improvement. Plotting the predicted vs. actually ob-
served proportions of full SRCs in Figure 5 shows a far from perfect fit. This is in part due to the low overall probability of full SRCs, which make it hard to accurately estimate effect sizes for bins for which almost no data is observed (i.e. anything above 0.15 on the x-axis of Figure 5). In the lower range of predicted probabilities, for which most data is observed, the model fits well.

![Figure 5](image_url)

**FIGURE 5** Mean predicted probabilities vs. observed proportions of full SRCs. The data is divided into bins based on 0.015 intervals of predicted values from 0 to 0.3 (there is not data with higher predicted or observed probabilities of full SRCs). The amount of observed data points in each bin is expressed as multiples of the minimum bin size. The top-most sunflower has 0 petals, so that the bin contains at least 50 data points. Each additional petal indicates 50 additional data points in that bin.

The only predictors that show signs of collinearity are the different components of the restricted cubic spline for SRC length. Since these components are only collinear with one another and since we are not interested here in their individual significances, these collinearities can safely be ignored (all other absolute fixed effect correlations < 0.2).

**Effect Evaluation**

As predicted by UID, both measures of the information density at the SRC onset have a highly significant positive effect on the log-odds of choosing the full form (I(SRC | NOUN): $\beta = 0.15, z = 3.7, p < 0.0002$; I(SRC | PARTICIPLE): $\beta = 0.77, z = 16.7, p \ll 0.0001$). The effects are visualized in Figure 6.

Most control parameters also had the expected effects. As predicted by availability-based production (V. S. Ferreira & Dell, 2000), the lower
FIGURE 6 Effect of information density on log-odds of SRC reduction. (a) shows SRC onset information content based on the preceding noun and (b) shows SRC onset information content based on the participle. For ease of presentation, visualizations are based on a model in which the information density measures were not residualized against the frequency measures (since residualization makes interpretation on the original scale difficult). Hexagons visualize the distribution of SRC onset information content against predicted log-odds of full SRCs considering all predictors in the model.

the log-transform frequency of the participle, the more likely producers are to produce the optional function words of full SRCs ($\beta = -0.43, z = -11.2, p \ll 0.0001$; all effects given in log-odds). For the least frequent participle in the database (146 occurrences in the written BNC), the odds of choosing a full SRC are 7.7 times higher than for the most frequent participle (85,136 occurrences). In contrast to a recent study I conducted on complementizer mentioning (Jaeger, 2010), there was no sign of ‘spill over’ effects: the log-transformed frequency of the modified NP’s head noun has no significant effect ($p > 0.8$). The effects of the two frequency predictors are illustrated in Figure 7.

There was no evidence for ambiguity avoidance accounts. Producers do not seem to be more likely to use the full form if the participle is ambiguous between a past tense and a passive participle interpretation ($p > 0.7$).

The effect of SRC length is also significant, as confirmed by comparison against a model without the restricted cubic spline of SRC length ($\chi^2_{\Delta(\lambda)}(3) = 52.6, p \ll 0.0001$). The effect, which is illustrated in Figure 8, contains a significant non-linear component ($\chi^2_{\Delta(\lambda)}(2) = 11.8, p \ll 0.003$).
3.6 General Discussion

Previous evidence for UID effects on syntactic reduction comes exclusively from that-mentioning (Jaeger, 2006, 2010; Levy & Jaeger, 2007; Wasow et al., in press). The current results extend this finding to another syntactic reduction environment, the reduction of passive subject-extracted relative clauses (SRCs). Unlike that-mentioning, SRC reduction
involves a chain of optional words that have to be produced or omitted together, thereby further illustrating that SRC reduction is best thought of as a choice point during syntactic production. Hence the evidence obtained here further corroborates the claim that syntactic production is affected by Uniform Information Density (Jaeger, 2006, 2010; Levy & Jaeger, 2007).

Both estimates of the information density at the SRC onset have highly significant independent effects on producers’ preferences. Information density is the single most important predictor of SRC reduction. Over 80% of the model improvement in terms of data likelihood compared to the base line model is due to information density (removal of information density measures from full model: $\chi^2_{\Delta(\Lambda)}(2) = 279.5$, $p \ll 0.0001$). A similarly strong effect of information density has been observed for complementizer that-mentioning (Jaeger, 2010), suggesting that a strong preference for uniform information density drives producers’ choices during syntactic production.

These results are encouraging for UID as a theory of sentence production. There is, however, also reason for caution. First, the results are based on data from a written corpus. Written text is subject to editing, which involves both comprehension processes and conscious reasoning about comprehension. It remains to be seen whether SRC reduction in spontaneous speech also follows the predictions of UID—as observed for that-mentioning in relative clauses (Jaeger, 2006) and complement clauses (Jaeger, 2010). Second, the model presented above requires improvement. Model evaluation suggested that the effect of some predictors is over- and/or under-estimated and that other important predictors may be lacking from the model. For example, the model is lacking controls for syntactic priming and social variables, both of which have been argued to affect other syntactic reduction phenomena (for an overview, see Jaeger, 2010). Fortunately, there is no obvious reason to expect these predictors to confound the effect of information density.

The analysis also provided evidence for availability accounts (V. S. Ferreira, 1996; V. S. Ferreira & Dell, 2000; Levelt & Maassen, 1981; Prat-Sala & Branigan, 2000). It is worth mentioning that the evidence in support of availability accounts, the negative effect of log-transformed frequency, can alternatively be interpreted as additional evidence for UID. Recall that the information density of the SRC onset is determined by at least two pieces of information: The information that an SRC constituent has started and the lexical information of the SRC onset. In this paper, I have focused on the former component, but UID predicts that producers should prefer full SRCs if either of these
two pieces of information is unexpected. Since a word’s frequency can be seen as a very simple estimate of its probability (ignoring context), and since the Shannon information of a word is defined as the negative log-transformed inverse of its probability, the negative effect of log-transformed frequency is compatible with the predicted positive effect of lexical information content at the SRC onset. Future work is necessary to test whether this frequency effect is due to availability-based production or UID (or both).

In line with previous findings on different syntactic reduction phenomena (e.g. Hawkins, 2001; Jaeger, 2006, 2010; Race & MacDonald, 2003; Tagliamonte & Smith, 2005; Tagliamonte, Smith, & Lawrence, 2005), the effect of SRC length provides support for dependency processing accounts (Hawkins, 1994, 2004, this issue). The non-linear effect of SRC length seen in Figure 8 above resembles that observed in previous work on that-mentioning in relative and complement clauses (Jaeger, 2006, 2010).

There was no evidence of ambiguity avoidance (contrary to Hawkins, 2004; Haywood et al., 2005; Temperley, 2003, but in line with V. S. Ferreira & Dell, 2000; Jaeger, 2006; Roland et al., 2005). It is possible that producers only avoid ambiguities that comprehenders are likely to stumble over (cf. Wasow & Arnold, 2003, Section 3.3; see also Arnold, this issue; Jaeger, 2006, 2010). In every day language use, context usually provides cues towards the intended parse. It is well-known that comprehenders are highly efficient in taking contextual cues into consideration (Trueswell, Tanenhaus, & Kello, 1993; Trueswell, 1996; Garnsey, Pearlmutter, Meyers, & Lotocky, 1997; Hare, McRae, & Elman, 2003; Hare et al., 2007; Wilson & Garnsey, 2009). This suggests that many—if not most—syntactic ambiguities are unlikely to cause garden paths. In other words, the lack of strong replicable ambiguity avoidance effects in corpus studies may be due to the rare need for ambiguity avoidance in contextualized spontaneous speech.6 Indeed, there is evidence from that-mentioning in complement clauses consistent with this hypothesis (Jaeger, 2010): where comprehenders would be likely to garden path and the ambiguity would be long lasting (i.e. when the disambiguation point occurs many words after the onset of the ambiguity), speakers do seem to be more likely to avoid syntactic ambiguity.

6This hypothesis would still leave a number of conflicting experimental results (cf. Haywood et al., 2005 vs. V. S. Ferreira & Dell, 2000; Kraljic & Brennan, 2005) unaccounted for. It is possible that the rare need for ambiguity avoidance may be the reason why ambiguity effects seem to depend a lot on the experimental items, task, and—potentially—individual differences between participants.
It is possible that the same holds for SRC reduction. Consider, for example, (6) where released is a potentially ambiguous verb form, but smell is a very unlikely agent to released. Comprehension experiments have shown that comprehenders are unlikely to be garden pathed in such environments (Trueswell, Tanenhaus, & Garnsey, 1994). Additional annotation will be necessary to determine whether producers avoid reduced SRCs if comprehenders are likely to be garden pathed otherwise.

(6) The smell released by a pig farm is indescribable.

In the current study, the ambiguity also lasted only briefly: The point of disambiguation for the SRCs was always immediately adjacent to the point of ambiguity since the participle was immediately followed by a by-PP. In short, for the SRCs in the current study, there would have been very little need for ambiguity avoidance.

Finally, note that audience design can be construed more broadly than ambiguity avoidance (see, e.g. Brennan & Williams, 1995; Clark & Carlson, 1982; Clark & Murphy, 1982). The effect of information density, for example, could be attributable to audience design (for discussion, see Jaeger, 2010). In particular, the fact that the strongest predictor of producers’ preference for a full SRC is the conditional (log) probability with which a participle occurs in a subject-extracted relative clause rather than in another syntactic environment (e.g. as a past tense matrix verb) is highly compatible with a more general audience design account. Consider that comprehension is a noisy process (e.g. because word recognition is not perfect), so that there are not just ambiguous vs. unambiguous cases, but rather more or less uncertainty about the parse. In that case, it is possible that producers aim to reduce uncertainty about the correct parse (rather than to avoid ambiguity). This interpretation of audience design would also predict the observed effect of information density. While this interpretation of the results is highly related to UID, it is not the same. I plan to explore this further in future work.

4 Conclusion

In conclusion, the corpus study presented above provides evidence that a preference for uniform information density drives producers’ choices in the syntactic reduction of passive subject-extracted relative clauses. This lends further support to the hypothesis that language production at all levels of linguistic processing is affected by information density (Aylett & Turk, 2004; Frank & Jaeger, 2008; Jaeger, 2006, 2010; Levy & Jaeger, 2007).
I have discussed many of the steps involved in corpus-based research in an effort to make this methodology more transparent to researchers entering the field. While corpus-based research requires complex modeling in order to deal with highly heterogeneous unbalanced data, corpus-based studies complement psycholinguistic experiments. The trade-offs of the two approaches are discussed in more detail in Jaeger (2010).

Acknowledgements
I am grateful to Benjamin Van Durme and Judith Degen for creating a parsed version of the British National Corpus. Special thanks also go to M. Dizon and C. Hansen-Karr for doing a superb job in leading the annotation team consisting of them and J. Ferris. Thanks for feedback on earlier versions of this paper go to J. Arnold, A. Perfors, S. Riehemann, K. Housel, C. Hansen-Karr, M. Dizon, C. Hilliard, and K. Gruters. This work was partially supported by NSF grant BCS-0845059 and by the 2008 Provost’s Multidisciplinary Award of the University of Rochester.

References
Arnold, J. E. (this volume). Ordering choices in production: For the speaker or for the listener? In E. M. Bender & J. E. Arnold (Eds.), *Language from a cognitive perspective: Grammar, usage, and processing*. Stanford: CSLI.


Hawkins, J. A. (this volume). Discontinuous dependencies in corpus selections: Particle verbs and their relevance for current issues in language processing. In E. M. Bender & J. E. Arnold (Eds.),
Language from a cognitive perspective: Grammar, usage, and processing. Stanford: CSLI.
Lohse, B. (2000). *Zero versus explicit marking in relative clauses.* (Department of Linguistics, University of Southern California)


Language from a Cognitive Perspective:
Grammar, Usage, and Processing
edited by Emily M. Bender and Jennifer E. Arnold

Published by CSLI Publications, Stanford,
and distributed by University of Chicago Press.

Information on buying the full book can be found at
http://cslipublications.stanford.edu/
or at
http://www.press.uchicago.edu/